DOCUMENT RESUME

ED 221 567 TM 820 590

AUTHOR Winters, Lynn; And Others
TITLE Bilingual Evaluation Technical Assistance: Workshop Text III. Planning for Student Assessment. Draft.
INSTITUTION California Univ., Los Angeles. Center for the Study of Evaluation.
PUB DATE Jun 80
NOTE 90p.; For related documents, see TM 820 588-592.

EDRS PRICE MF01/PC04 Plus Postage.
DESCRIPTORS *Bilingual Education; *Bilingual Students; Educational Planning; Elementary Secondary Education; *Evaluation Methods; Evaluation Needs; Program Evaluation; *Technical Assistance; Test Bias; Test Reliability; *Test Selection; Test Validity; *Workshops

ABSTRACT

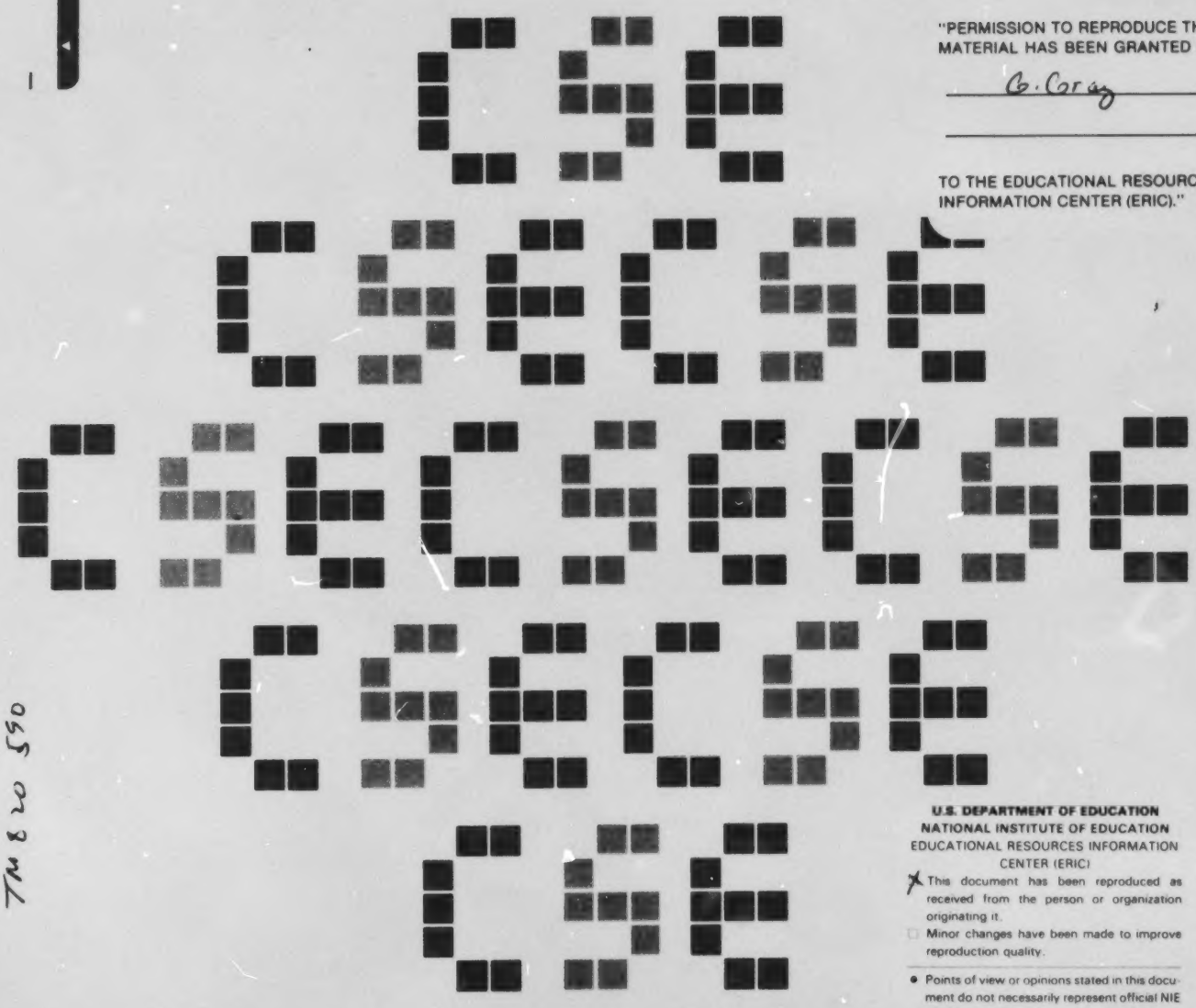
This text is designed to accompany workshop instruction in student assessment technique for the purpose of evaluating bilingual programs. It is addressed to those involved with bilingual programs generally, educators directly involved in planning and implementation, and research and evaluation specialists. The contextual variables that affect the selection of appropriate student assessment instruments are summarized as a guide. Screening criteria (validity, reliability and practicality of test use) are discussed in the context of the program being evaluated, the program demands, and the constraints of data collection for bilingual programs. Managing the collection and reporting of test data, and features to consider when selecting a test for bilingual program evaluation are appended. (PN)

**********
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
**********

**Center for the Study of Evaluation**

UCLA Graduate School of Education
Los Angeles, California 90024

ED221567



"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

G. Coray

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 820 590

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

Bilingual Evaluation Technical Assistance Workshop Text ~~IV~~ III:

# PLANNING FOR STUDENT ASSESSMENT

Developed by:

Lynn Winters

Beverly Cabello

Herminio Martinez, Columbia University

Marietta Saravia Shore, Columbia University

A special thank you to our reviewers: Elois Scott, Frank Capell and Joan Herman

Bilingual Evaluation Technical Assistance

Workshop Text III

- Overview: Evaluation of Bilingual Programs
- Planning for Evaluation Design
- Planning for Student Assessment

## How to Use This Text

Planning for Student Assessment is designed to accompany workshop instruction in evaluating bilingual programs. Instructional activities to provide participants with practice on skills presented in this workshop can be found in Participant's Handbook III: Planning for Student Assessment.

You will want to read this workshop text before participating in the actual sessions. You might want to complete the "Participant Preparation" form provided at the beginning of the Participant Handbook accompanying this text. Bring questions you have and examples from your own program to the workshop.

Planning for Student Assessment addresses three levels of involvement with bilingual evaluation. At the most general level, we hope that all people involved in bilingual program, both bilingual and non-bilingual program teachers, students, parents, and community members will find the materials valuable. A second audience is educators directly involved in

bilingual program planning and implementation who are familiar with many of the general principles of curriculum development and testing, but who have no special training in evaluation. The third group for which the materials were designed is research and evaluation specialists who may not be well-acquainted with the special demands of bilingual program evaluation.

Since sections of the test will be more relevant to some groups than to others, the Table of Contents is keyed by section. Consult the key to identify those parts of the test that match your interests and level of expertise.

Table of Contents


| | Intended Audience | Page |
|---|---|---|
| Part I | | |
| Introduction. . . . . | | 1 |
| Testing Non-Native English Speaking Students . . . . . | | 2 |
| Finding Appropriate Instruments . . . . . | | 4 |
| Language Assessment in Bilingual Programs | | |
| Assessment Requirements . . . . . | | 6 |
| Terminology in Language Assessment . . . . . | | 6 |
| Language dominance . . . . . | | 6 |
| Language proficiency . . . . . | | 7 |
| Communicative competence . . . . . | | 7 |
| Language achievement . . . . . | | 7 |
| Language test tasks . . . . . | | 8 |
| Test content . . . . . | | 9 |
| Avoiding Overtesting in Bilingual Programs . . . . . | | 12 |
| Issues and Problems in Multilingual Testing . . . . . | | 13 |
| Test misuse disguised as bias . . . . . | | 14 |
| Construct bias . . . . . | | 14 |
| Content bias . . . . . | | 15 |
| Finding Language and Achievement Tests for Your Program . . . . . | | 16 |

Table of Contents(cont)


| | Intended Audience | Page |
|---|---|---|
| Part II | | |
| Selecting Tests for Bilingual Programs | G/S | 18 |
| Types of Tests | | 19 |
| Standardized tests | | 19 |
| Criterion Referenced and Norm-Referenced Tests | | 20 |
| Norm-referenced tests: Construction Procedures | | 21 |
| Norm-referenced tests: Score interpretation and use | | 22 |
| Norm-referenced tests: Limitations | | 22 |
| Criterion-referenced tests: Construction | | 23 |
| Criterion-referenced tests: Score interpretation and use | | 24 |
| Criterion-referenced tests: Limitations | | 24 |
| Minimum-Competency Tests | | 26 |
| Minimum-competency tests: Limitations | | 27 |
| Part III | | |
| Technical Quality | S/R | 29 |
| Validity | | 29 |
| Threats to validity | | 30 |
| A note about out of level testing | | 32 |
| Guidelines for selecting valid tests | | 32 |

Table of Contents(cont)


| | Intended Audience | Page |
|---|---|---|
| Reliability | | 33 |
| Split half reliability | | 35 |
| Guidelines for selecting reliable tests | | 35 |
| Practicality | | 37 |
| Administrator qualifications and training | | 37 |
| Test directions | | 37 |
| Test formats | | 38 |
| Scoring | | 38 |
| Cost | | 40 |
| Summary | | 41 |
| Appendix A | | |
| Managing the Collection and Reporting of Test Data . R | | 44 |
| Preparation for Data Collection | | 44 |
| Important Datas in Scheduling Data Collection | | 49 |
| Recording Data for Analysis: General Considerations | | 50 |
| Determining what to record | | 50 |
| Individual item responses | | 52 |
| Item scores | | 52 |
| Raw v. derived scores | | 52 |
| Recording Systems | | 55 |
| Individual student data record form | | 55 |
| Classroom data record form | | 55 |

Table of Contents(cont)


| | Intended Audience | Page |
|---|---|---|
| Program data record form | | 57 |
| Reporting Test Results | | 63 |
| What do NRT Printouts Tell Us? | | 65 |
| What do the Scores Mean? | | 66 |
| Recommended Scores | | 67 |
| Scale scores | | 67 |
| Scores Not Recommended | | 68 |
| Percentiles | | 68 |
| Grade equivalents | | 69 |
| How Do I Interpret These Scores | | 70 |
| Norm group interpretations | | 70 |
| Curricular Implications of Scores | | 70 |
| Score accuracy | | 71 |
| Appendix B | | |
| Thirty-three Features to Consider When Selecting a Test For a Bilingual Program | | 74 |

Bilingual Evaluation Technical Workshop

Workshop Text III

## Introduction

Everyday in classrooms throughout the nation, some form of student assessment takes place. Teachers make informal judgments about their students' progress in order to provide appropriate instruction. More formal testing also occurs for assignment of students to special programs or for diagnostic-prescriptive instruction. Students also take tests to meet legislative and funding agency requirements.

Given that testing is a common practice in the schools, why devote more pages to a topic already thoroughly discussed, analyzed, and practiced? Bilingual programs, after all, have much in common with other educational programs. Like most educational programs, bilingual instruction is characterized by:

- a diversity of goals
- competing theories of instruction
- a heterogeneous student population

What distinguishes the assessment activities of bilingual programs from those in other settings are two factors:

- tests are used to make decisions about students whose native language is not English
- appropriate instruments for this purpose are generally unavailable

Testing Non-Native English-Speaking Students

The problem of testing students for whom English is a second or third language is perhaps the major issue when planning for student assessment in bilingual programs. Because the "test language" is not the student's native language, it is difficult to determine whether test performance is a function of language proficiency or school achievement in the target area. It is unfortunate that this "problem", the confounding of language proficiency with achievement, has been associated particularly with bilingual programs. The potential for confounding achievement estimates with estimates of language proficiency exists in all formal testing programs for all students. The interrelationship between language and cognitive development, so important in the assessment of bilingual students, holds too for English monolinguals. Instruction and assessment are language-based activities in all educational programs; thus, the "problems" in language proficiency assessment and achievement testing faced by bilingual evaluators are but a special subset of the kinds of assessment problems found in all evaluations.

What makes student assessment in bilingual programs unique is the necessity to provide <u>separate</u> data about language and cognitive development. Most funders require information about students' "Language proficiency" and "achievement in basic skills." Unlike monolingual programs, bilingual assessment information must be gathered to document <u>both</u> language and cognitive development. In other programs, only achievement (cognitive development) is assessed, although most participants recognize that academic growth is closely tied with language proficiency.

Because of the necessity to separate language proficiency and achievement data, bilingual measurement specialists have had to tackle the issue: What is language proficiency and how can it be measured? Answers to these questions vary. Bilingual measurement specialists differ in their approaches to language assessment. For example, some experts use the terms "dominance" and "proficiency" to refer to different kinds of language tests and decisions about students. They hold that these terms have meaning for instructional planning. Others feel that the concept of proficiency is the only one with utility for educational decisions and that testing for placement and progress should be used to estimate language development. Some advocate testing student mastery of particular grammatical structures and vocabulary; that is discrete point testing. Others say that language proficiency is best estimated from language production tasks such as speaking, writing or taking dictation, called integrative testing. Even within these two groups there is controversy. For those in favor of objective/discrete point tests, some would suggest that English tests can be translated to assess similar concepts in the native language. Their adversaries argue that content, skills and concepts identified as appropriate subject matter for English-speaking students at certain developmental levels, may not be for students from other cultures. Translations, they hold, do not have the same level of difficulty as the original test. Among those who favor integrative tests, there is debate as to the amount of contextualization necessary for the "test task" and how generalizable test results can be.

The differences among bilingual testing experts about approaches to

language testing remain unresolved. Consequently, we do not recommend particular language assessment strategies. Instead, we present a description of the terminology and assessment instruments currently used. Your position on language assessment will be determined by the region of the country in which you are located, legislative requirements, your own experience and the kinds of evidence you can gather about a test's validity for your program.

Finding Appropriate Instruments

The second assessment issue in bilingual programs, that of a lack of technically sound and curricularly-relevant instruments for your program, is somewhat more tractable. In the past decade, educators have begun to use tests to supplement information obtained from nationally-standardized, norm-referenced tests. Minimal competency tests are an example of this trend. The meaning of the term "assessment" has broadened to include many other systematic data collection techniques, such as observations, attitude surveys, questionnaires, interviews, and archival data. These alternative assessment strategies are often used in conjunction with achievement test results to clarify, support, or explain program effects. A convergence of data from multiple measurement strategies is often more convincing and informative than single score reports. Thus, the problem of inadequate instrumentation, while not solved, can be blunted with multiple measures or alternative strategies assessment plan. Alternative assessment strategies are described in detail in BETA Workshop IV: Planning for Implementation Evaluation. You will want to refer to them after reading this text.

Although "outcomes" remain a primary interest of many evaluations, "outcome" is not synonymous with standardized achievement test scores alone. Programs have many outcomes, affective, cognitive, and social, that may promote student achievement and be considered equally important in themselves. Although discussion in this text does focus on test data, your evaluation will include other important outcomes as well. The instruments selected to assess different kinds of outcomes will vary according to what is being measured and may look quite different from the standardized, paper and pencil instruments described herein. Nevertheless, these alternative assessment strategies will need to meet the same criteria for validity and reliability presented in this workshop, thus making a review of test selection criteria worthwhile.

The purpose of Planning for Student Assessment is to introduce you to:

1. Language assessment concerns in bilingual programs and
2. Test selection criteria for reviewing language and achievement tests.

For those of you to whom the problem of gathering and recording test data is new, you will find a general set of procedures for managing the collection and reporting of test data appended. We have not attempted to provide model reporting plans for bilingual programs, as these will vary according to project requirements. Our goal throughout this text has been to present information about testing common to a majority of bilingual evaluation efforts that can be adapted to the assessment needs of your particular project.

## Language Assessment in Bilingual Programs

### Assessment Requirements

Since the 1974 Lau v. Nichols Supreme Court decision, schools are required to identify and provide appropriate educational services to students whose language proficiency interferes with their ability to learn in the regular educational program. In order to comply with this decision, schools must determine whether a student's cognitive and linguistic development would be promoted to a greater extent through placement in a bilingual than a monolingual program. In order to provide these equal educational opportunities, schools must determine which students are unable to achieve satisfactorily in a regular program and then ascertain whether poor achievement is due to a lack of English language proficiency. Assessment in bilingual programs, then, is generally of two sorts: language testing and achievement testing. Since bilingual program placement, progress and exit is determined primarily by a students' language proficiency, discussion in this section will focus on language assessment. Issues to consider in achievement testing will be discussed at the end of this section in "Issues and Problems in Multilingual Testing."

### Terminology in Language Assessment

We begin our discussion of language assessment with a clarification of the terminology used to describe the different kinds of language tests.

Language dominance. Language dominance is often used to refer to the language that the student is most comfortable using. Often, dominance is situationally determined; a student may use one language in the home

and the other in the school with little interference. An example of this would be a child who knows the names of foods and kitchen instruments in his/her native language, and of school objects in English, yet would be very unlikely to use English in the kitchen or the native language in school.

Language proficiency. Language proficiency refers to the degree to which an individual demonstrates his/her linguistic competence in a language regardless of how he/she acquired that language. There are different components of language proficiency. Basically, language proficiency includes oral production, or speaking skills, and listening comprehension skills. Reading and writing proficiencies are also considered as a component of language proficiency. Some educators believe that communicative competence is another component of language proficiency. Communicative competence is the knowledge of when to speak and when to remain silent; what to say to whom and when (Hymes, 1972). Language proficiency tests may measure either linguistic competence, communicative competence or both. The criterion for competence is generally the standard adult language in a given speech community. Tests of student language proficiency, then, measure how much a child's language deviates from the adult model.

Language achievement. Achievement tests measure a student's learning history, that is, acquisition of course content. Language achievement tests measure these aspects of phonology, lexicon, and structure to which the student has been formally exposed in textbooks, classroom session, or through other instructional means. For the child exposed to standard English, an English "achievement test" may be similar to a proficiency

test, for it is a test of naturally and developmentally acquired language. For the foreign or non-standard English speaking child, language proficiency tests that focus on standard English grammar, vocabulary and syntax may be achievement tests because school is their source of language acquisition. Achievement tests are generally of two sorts: general and diagnostic. General achievement tests measure a student's ability to combine several different aspects of course content in situations which closely approximate ordinary language use. Diagnostic tests measure discrete aspects of course content.

The difficulty in labelling a test an "achievement" test stems from the fact that once a child enters school, it is almost impossible to specify exactly which knowledge is attributable to course content and which is acquired somewhere along the way. For some students whose instructional history differs from the group being assessed, an achievement test is more like a proficiency or ability test. The same may be said for children who were "instructed" but did not "learn." While most achievement tests purport to measure school-based skills and content, whether in fact they do is as much a function of the kinds and quality of instruction provided as of the test itself.

Language test tasks. A student's mastery of English structures may be tested in many ways. Language acquisition has been assessed at the secondary level, by having the student translate passages. This method taps lexicon and comprehension to some extent, but is difficult to score reliably because of the problem of judging tone and mood in the translated piece. Dictation measures phonological, grammatical and lexical competence, but is terribly uneconomical in terms of scoring time, and it can be imprecise. Student

compositions can be used to assess grammar, lexicon, and the student's grasp of graphic conventions and organization. However, the difficulty of generating appropriate writing prompts for eliciting the structures you wish to test and of finding reliable, economical scoring systems is amply documented. Scored interviews do provide a valid measure of student language proficiency, but they are difficult to standardize in terms of prompting questions. Unless the interview is taped, rater reliability is usually difficult to establish. Multiple choice questions provide the most frequently-used method of achievement testing. These kinds of tests take much time and skill to construct and do not always satisfy the evaluator's criteria for construct validity. Finally, short answer questions are often used in assessing student achievement, especially in the classroom. These questions are easy to write and take less time to score than full-blown compositions. Developing scoring criteria that anticipate the range of possible answers (correct) that students might produce is often impossible. Thus, each method of assessing language proficiency/achievement has its limitations.

Test content. The kinds of content that can appear on language proficiency and achievement tests, and some of the difficulties in assessing this content was carefully presented by Harris (1969). Harris suggested that six areas of language should be assessed: grammatical structures, auditory discrimination and comprehension, vocabulary, reading comprehension, writing and oral production. Each of these language areas presents certain problems for test developers; a review of these problems should alert you to possible deficiencies in language tests that you will select.

The intent of measuring a student's grasp of grammatical structure is to determine his/her control of basic spoken grammatical patterns. Most

high school level English achievement tests measure formal writing style and diction and are therefore not appropriate for measuring the English language skill of all but the most advanced of foreign students. What should be included on tests of grammatical structure is a broad range of grammatical problems that non-native speakers encounter in speaking English. These "problems" should be represented in terms of their relative importance in learning the language, rather than structures that are easy to test.

Auditory discrimination and comprehension usually is measured by determining how well students respond to one sentence requests, questions, statements of fact, conversations, or oral discourse. Tests of auditory discrimination, unless administered by an invariant script or tape recorder, suffer serious problems of score reliability. Unless the test prompts are standardized and scored by a stable rating system, measures of auditory skill will be unreliable.

There are many different opinions as to what constitutes a good source for the vocabulary tested on English proficiency/achievement tests. One of the first decisions you will make in selecting an appropriate vocabulary test is whether you wish to test "active language," speaking and writing, or "passive language," reading and recognition, vocabulary. Harris suggested that no matter what the source of the vocabulary included on the test, idioms and two word verbs such as "look over," "put off," are important indicators of English proficiency.

At least three aspects of reading comprehension might be included on an English achievement measure for non-native speakers, language and graphic symbols, ideas, and tone. Reading comprehension should measure several aspects of the students' ability to infer meaning from passages,

from picking out the main idea, to identifying support for a thesis, to drawing conclusions. Measures of a student's comprehension of tone and style include the author's attitude toward the subject and knowledge of stylistic devices. A reasonable average length for comprehension passages is from 150-200 words with internal organization which is either chronological, comparison and contrast, or a presentation of an author's opinions on a familiar subject. In examining comprehension items, you should check to see that they cannot be answered correctly without reference to the passage, i.e., that these are passage dependent. Ideally, reading comprehension tests will not require students to have mastered outside subject matter in order to understand the passage, thereby giving a marked advantage to persons in particular fields.

Writing is a skill that is best measured by asking students to produce an essay or writing sample. Problems encountered in assessing writing skill include selecting a topic that is familiar enough to a wide range of students and which enable them to be able to write themes long enough for a reader to assess their skill in organization, diction, support, mechanics. Clear specification of criteria for scoring is another problem. Different raters judging the same theme should be able to reach the same decision about the adequacy of the writing sample. (These problems are dealt with in greater detail in Thorndike, 1971.)

The assessment of oral production has difficulties analogous to the measurement of writing skill. There is a problem in specifying what constitutes "good" or acceptable production and adequate responses. Techniques to improve the reliability of writing measures apply also to the measurement of oral production. These include: 1) the use of two or more raters to judge the oral presentation; 2) a clear specification of methods and criteria for judging the adequacy of responses.

## Avoiding Overtesting in Bilingual Programs

The previous discussion about the different kinds of language and proficiency tests that might be used to help make decisions about bilingual students for placement, progress, exit and evaluation may make it appear that there is a tremendous amount of testing taking place in these programs. While this can be the case, test results can be used for several purposes, thereby cutting down on the amount of required testing. In order to minimize testing, and to identify those tests that can be used for multiple purposes, you may wish to map the testing requirement of your program. For example, a comprehensive diagnostic tool used for language proficiency assessment can also provide information for program placement, diagnosis for instruction, and progress monitoring. Or, you could use state-mandated achievement test results to report required achievement information to funding agencies. You might also consider reviewing student test data collected in other classes/programs for progress monitoring, exit testing, or evaluation. Figure 3 below provides an example of what a map of testing requirements and test would look like to help you identify areas where test results can be sued for multiple purposes.

Figure 3-1

Instructional Programatic Decisions

| Types of Assessment | | Factors Being Assessed | Language of Measurement |
|---|---|---|---|
| Language Assessment | 1. Entry<br>2. Instruct. Place.<br>3. Exit | Listening. Comp.<br>Speaking Fluency<br>Reading Compre.<br>Writing Compre. | L1 and L2 |
| Achievement | 1. Monitoring Sch. Learning<br>2. Program's Eval. | Lang. Arts<br>Math<br>Social Stud.<br>Sciences | Language in which instruction is given |
| Special Bilin. Educa. Programs | 1. Psychological Diagnoses<br>2. Instructional Placement | IQ Tests | Students' strongest language |

Issues and Problems in Multilingual Testing

The assessment of language proficiency, achievement and the evaluation of the impact of bilingual educational programs is limited by the availability of appropriate assessment instruments in L1 and L2. Several issues are pertinent to the selection, use, and/or development of tests in languages other than English. Some of these issues are specific enough to be treated in the context of bilingual evaluation. Other concerns are so abstract and general that their determination could only be made on theoretical grounds. Depending on the language in which the testing is to be done, these issues take different pragmatic characteristics. These issues are:

1. The use of translated/adapted tests from English to the other language;
2. The construction of tests in English and another language so that they are equivalent in their content as well as in their psychometric properties;
3. The use of different tests for assessing skills in each language, which produces results which cannot be compared across languages;
4. The use of tests which have been developed and normed for a student population different that the population being assessed;
5. Availability of appropriate assessment instruments;
6. Availability of bilingual personnel either in the school districts or in other institutions, with skills needed for developing bilingual tests.

Most of these problems are generally clustered under the label of "test bias".

Test bias is defined as the differential performance of subpopulations on tests that is not attributable to instructional history. Examples of bias most often occur when members of linguistic or ethnic minority groups are found to score systematically lower on tests than members of mainstream groups. The issue is, "Why do these students have differential performance patterns?" Is it due to instructional experiences, lack of test taking skill or is there something in the content of the test that "causes" students to answer questions in an unexpected way? The potential for systematically differential student performance on tests is especially great in bilingual programs, where students are of diverse backgrounds. Existence of these different patterns of test scores, of course, does not necessarily mean that the test itself was biased; the evaluator may have selected an inappropriate test. There are three general explanations for "biased" test results; one stems from test-examinee mismatch, one from a mismatch between the assumptions of the test developer and that of the examinee and one from poorly selected content.

Test misuse disguised as bias. Misuse of tests occurs when incongruities arise between the assumptions underlying test development and the setting in which the test is being used. For example, if you select a "Spanish" translation of a test laden with vocabulary and content typical of Caribbean cultures and use it with Mexican-American students, it will be "biased" against that group. To ensure a test is appropriate for your students, you will need to have the test reviewed by staff members who are familiar with both the curriculum and cultural background of the examinees.

Construct bias. A test that measures different aspects of achievement

in different populations may have construct bias. That is, it really is measuring different traits in different populations. For example, a reading comprehension test may measure these skills for one group but scores for another group may reflect "background" or vocabulary knowledge only. The presence of construct bias is detectable when the factor structure of the test for differing examinee populations is examined. Significant differences in the pattern of factor loadings for the two groups or significant deviations from a prespecified common factor structure may indicate that the test does indeed tap separate constructs in different subpopulations. Tests that assess different constructs in different subpopulations are invalid for making decisions about the group for whom the test is not functioning as intended, or for making comparisons between the groups.

Content bias. The content of a test may be inappropriate for subgroups of students for several reasons: 1) there is sex-role stereotyping in the content of the test; 2) test content refers to material usually learned in out-of-school settings and thus is biased against students from non-mainstream backgrounds; 3) test vocabulary may have culturally determined meanings; 4) there are discrepancies between test objectives and curricular goals. Content bias is detected by direct examination of the test items in relationship to what is being taught and the students' socio-economic, linguistic and cultural backgrounds. If you are concerned that the tests selected for your evaluation may be biased, turn to Appendix B, Thirty Three Features to Consider When Selecting a Test for a Bilingual Program. The guidelines presented in this section will help you to avoid

test misuse, to identify poorly constructed tests and to determine if the assumptions of test development match the assumptions underlying your bilingual curriculum.

Finding Language and Achievement Tests for Your Program

The discussion of language testing presented in this section highlighted general functions and features of language and achievement tests for bilingual programs. It also mentioned some pitfalls to be avoided in multilingual testing. The availability of suitable instruments for bilingual evaluation is limited. There are, however, several catalogues of bilingual tests that will provide you with detailed information about particular language and achievement tests that will enable you to make a preliminary selection of tests. These catalogues are listed in Table 1. After referring to a catalogue and requesting examination copies of tests that you wish to consider, you will then need to review them according to criteria that will enable you to choose a high quality and practical measure. The following section, "Test Selection," provides guidelines for the review process.

Table 1

Catalogue of Language Tests

Backer, Thomas E. 1977. A Directory of Information on Tests. ERIC Clearinghouse on Tests, Measurement and Evaluation. TM Report 62.

Bye, Thomas T. 1977. Tests that Measure Language Ability. BABEL/LAU Center, Berkeley Unified School District.

California Association for Bilingual Education. 1976. A Qualitative Review of Instruments used in Bilingual Education Programs in California. Test Evaluation Conference, San Jose, California, April 1976.

Center for Bilingual Education, Northwest Regional Educational Laboratory. 1978. Assessment Instruments in Bilingual Education. National Dissemination and Assessment Center. Los Angeles.

Grommon, Alfred H., ed. 1976. Reviews of Selected Published Tests in English. Urbana: National Council of Teachers of English.

Locks, Nancy, Pletcher, B., & Reynolds, D. 1978. Language Assessment Instruments for Limited-English-Speaking Students: A Needs Analysis. National Institute of Education.

Oller, John W. & Perkins, K. 1978. Language in Education: Testing the Tests. Rowley, Mass.: Newbury House Publishers.

Pletcher, Barbara, Locks, N., Reynolds, D. & Sisson, B. 1978. A Guide to Assessment Instruments for Limited-English-Speaking Students. New York: Santillana Publishing Company.

Savard, Jean-Guy. 1977. Analytical Bibliography of Language Tests. Quebec: Les Presses de l'Universite Laval.

Silverman, Robert J., Noa, J. & Russell, R. 1976. Oral Language Tests for Bilingual Students: An Evaluation of Language Dominance and Proficiency Instruments. Portland, Oregon: Northwest Regional Educational Laboratory.

## Selecting Tests for Bilingual Programs

When selecting tests to meet the assessment needs of bilingual programs, you will need to review them with three questions in mind:

What type of test will provide the information required?

Is the test of good technical quality?

Do utility and cost factors make the test practical to use?

The purpose of this section is to provide guidelines for reviewing different types of tests, their technical quality and their usefulness. Each of these characteristics of a test contributes to or detracts from its appropriateness for evaluation. Since most tests vary on each of these dimensions, you will need to decide what kinds of compromises can be made in the information gathered, technical aspects and utility and still gather credible data. By assessing the strengths and weaknesses of the tests you use, you will also be able to decide what other kinds of information, from student work samples to archival data (attendance, referrals) you will need to support inferences about student achievement and clarify the implications of test score reports.

The discussion in this section is targeted to bilingual programs in general. Since language assessment issues, what to test, how, the use of translated tests, construct validity, are still being investigated in the context of bilingual evaluation, no attempt is made to recommend particular types of tests, statistical properties or testing procedures. You will also notice that we refer here to achievement tests. The criteria for reviewing observations, questionnaires, and other systematic data collection procedures are essentially the same as for standardized achievement measures. These alternative assessment procedures are discussed in BETA Workshop IV: Planning for Implementation Evaluation.

## Types of Tests

Naturally, extreme care must be taken in reviewing and selecting measures for bilingual students. Achievement measures are intended to measure what skills students have learned in school. Students cannot demonstrate these skills unless they comprehend test directions and test content. The validity and fairness of achievement testing is thus confounded with language skills. Available instruments should be reviewed and interpreted with this issue in mind.

### Standardized Tests

One category of tests you will review is standardized tests. Standardized tests have the following characteristics: their test items were written to prespecified content guidelines. The items were tried out on students and selected for inclusion on the test according to explicit criteria. The test was then tried out with a standardization sample of students in order to obtain reliability and validity statistics. These tests are also accompanied by a manual explaining administration, scoring and score interpretation procedures. Standardized tests include any kinds of assessment instruments fitting the above criteria. They may be produced by a test publisher, state of local education agency, or even an individual. These tests stand in contrast to teacher-made classroom tests or informal interview, observation and reporting techniques.

Standardized tests are designed to give you two kinds of information: information as free as possible from threats to validity posed by different examination procedures and scoring criteria; and information of a comparative nature. The standardization procedures are used precisely because they control from many of the competing explanations for different examinee results such as, "the students in group X had clearer directions," or "papers in group Y were graded more leniently." Not all of your evaluation information will come from standardized tests, but data used to make comparisons between groups should be gathered through standardized procedures.

## Criterion-Referenced and Norm-Referenced Tests

Although standardized achievement tests may appear to be similar in format, they may be quite different with respect to the types of interpretations they permit. Specifically, some achievement tests yield norm-referenced interpretations, (i.e. they tell how a student's performance compares with others) others yield criterion-referenced interpretations, (i.e., they tell what skills students have acquired). The distinction between these two types of achievement tests is extremely important and has direct implications for your evaluation.

Norm-referenced and criterion-referenced score interpretations are typically produced by different types of tests. It is possible that a test can be designed to yield both types of information; however it is more usual that achievement tests are designed for one purpose or the other. The different strategies used to develop norm and criterion-referenced tests highlight clearly the underlying differences between the two test types.

Norm-referenced tests: Construction procedures. As with standardized tests, norm-referenced test construction starts with the specification of test content. Broad objectives or content areas are selected and a large number of items are written to assess each. Items are tried out with a group of examinees selected to represent the kinds of students for whom the test is designed.

Data from this initial try-out is critical for developing a test that will spread out student scores in a normal distribution. Several item statistics are computed to achieve this end. One statistic, the item difficulty index, is used to delete items that too many students pass or fail and to retain items that are answered correctly by about half of the sample group. Another statistic, the item discrimination index, is used to locate and delete test items which do not discriminate between high and low achievers. In addition, in order to locate problems in multiple choice items, the number of students answering specific response alternatives may also be inspected. Once "good" items have been identified through the use of the item statistics, test forms are constructed and alternative forms of the test tried out again with an appropriate sample of students.

On this second tryout, the test developer is interested especially in norming the test and obtaining estimates of test reliability and validity. The performance of students participating in this norming tryout become the comparison group for interpreting test scores. For example, if three percent of the norm sample got four items correct on a 100 item test, the norm tables would say that a score of four would be interpreted as falling in the 3rd percentile. Because student performance on a norm-referenced test derives meaning principally by comparison to the norm group, it is critical that the norming sample be representative of the students for whom the test is designed. The ideal norm group would represent the student population on the following dimensions: geographic region of the

country, school size, community size, student demographic characteristics, e.g., socioeconomic status, ethnicity, language, school staff and instructional materials characteristics, and school type (public and private). Because norms become outdated, tests are usually renormed every three to five years.

Norm-referenced tests: Score interpretation and use. The interpretation of norm-reference scores emerges directly from the test construction procedures, i.e., selection of items that spread group performance into a normal distribution. Test scores provide information on how students compare to a national or local group of students who have taken the same tests, and are expressed as percentile, stanine, or standard scores which have been derived from the raw scores. Because test scores tell how well a student or group performance compares with the peer group standard, norm-referenced tests are suitable for making two kinds of decisions:

> Who shall be selected into a program when there are only a limited number of spaces available?
>
> How do our students compare with students of similar age and background on this skill area?

Norm-referenced tests: Limitations. There are some limitations on the use of norm-referenced scores that must be recognized. Because norm-referenced tests are designed to decide how well a student's performance in a subject area compares with other students, they may yield limited information about the amount of specific skills a student has mastered. Subscale scores are sometimes not sufficiently reliable, and often do not include a sufficient number of items to tell you if a student has mastered a particular objective. Test construction procedures which emphasize item selection on the basis of "appropriate difficulty level"

(.4-.6 difficulty level) and discrimination between high and low achievers may result in the inclusion of tricky or ambiguous items that relate little to content mastery or school instruction.

Criterion-referenced tests: Construction. Because the emphasis in criterion-referenced testing is on content as opposed to comparison, item construction procedures focus on selection of content for the test. Test construction begins with the specification of the concrete objectives to be tested; ideally detailed content specifications are designed to include all of the salient parameters of content or skill domain, such as eligible content, vocabularly, syntax, readibility, how distractors are to be written, etc. Based on the objective specification, teachers should be able to know exactly what is being tested. Based on these detailed objectives or domains, many items are then generated to reflect each objective. Like norm-referenced tests, these items are subsequently also tried out on appropriate samples of students. However, the purpose of the tryout is not to "weed out" items that do not discriminate between high and low scorers, but rather to identify items that appear to be measuring the same objectives and eliminate items that do not. Items selected for criterion-referenced tests in principle are selected for their ability to represent the content rather than to separate students. The ultimate criterion for item selection, then, is the relevance of the item to the objective being tested. Often items that might be discarded from a norm-referenced test because they are "too easy" or "too difficult" are retained in a criterion referenced tests because they measure important skills.

Like norm-referenced tests, the best items are then assigned to test forms and administered to representative groups of students. On the basis of these results, test reliability and validity is determined.

The results of students performance during this final try-out may also be reported in the test manual to give users some basis for comparing their students' scores with other students, although providing comparative data is usually not an area of emphasis.

Criterion-referenced tests: Score interpretation and use. As a result of these test construction procedures criterion-referenced test scores tell you how much of the content a student has mastered. Scores are often reported in terms of the number or percentage of items correct for each objective. Scores tell you about how well students have mastered the basic concepts described in the test specifications; they usually do not tell you if this compares well or poorly with the performance of other students in the same grade (although, some test developers do provide you with information about the performance of a comparison group). Criterion-referenced interpretations are especially appropriate for the following kinds of decisions:

How well have students mastered program objectives?

What are individual students' strengths and weaknesses in this area?

Which objectives am I teaching effectively and which not so well?

Criterion-referenced tests: Limitations. Criterion-referenced tests, too, suffer from some disadvantages. Because these are relatively new kinds of achievement measures, the technical quality of many of these tests is questionable, technologies for determining test quality, in fact, are still emerging. Although these tests purport to give you specific information about student performance, many tests do not contain a sufficient number of items per objective to be definitive (five items per

objective is a recommended minimum). Further, test specifications often are not as fully described or as well-defined as they should be. In addition, test objectives are often narrowly focused; as a result, a single test is not equally suited to the curricula of different schools or classes, making inter-school or program comparisons difficult.

Using Standardized Tests for Evaluation of Bilingual Programs. When deciding whether to use a NRT or CRT for bilingual evaluation, several issues should be considered.

Some educators argue against using standardized tests at all for evaluating bilingual programs because they feel it is unfair to judge students in these programs against students who are monolingual English speakers and whose culture is that of "mainstream" American society. This judgement is based on concerns about test bias and about the special skills required to take these tests that may be culturally specific.

Students in bilingual programs have the double task of mastering two languages and two cultures simultaneously. This may affect their learning rate; consequently, their test performance. If student achievement scores for bilingual programs appear to be low, the lower scores may be due to these two factors. First, students in bilingual programs may appear to progress at a slower rate because they have twice the amount of material to cover, language proficiency and content. Second, because most students in bilingual programs are not acculturated, they may be unfamiliar with test taking techniques as well as test content.

There are, of course, NRTs that have been normed with samples of different ethnic backgrounds and that have been analyzed for linguistic and cultural bias. When deciding whether to use an NRT or CRT, you should review it for cultural and linguistic appropriateness.

Whether bilingual students should be compared with monolingual students depends on the program's standard for success. State and federal government require comparative data. This preference for comparative data does not mean an NRT must be used. If you use a CRT, the standard for measuring student success will be a pre-specified body of knowledge. If you use a NRT, then your standard is the performance of the norm group. If one of your goals is to have students performing like a norm group, you will need normative data; but if your goals are to have students master pre-specified objectives, you will want CRT results.

One of the strongest arguments for using a CRT in your evaluation is its curricular relevance. While there are many NRTs which are objectives-based, the actual description of the domains covered in the test is not very specific and, hence, are quite limited for instructional planning or diagnosis.

In making your selection decision, it is important to remember that no single type of test score gives all of the needed information about students' performance. It is useful to know both how common or rare a given score is (which is NRT information) and what that score means in terms of mastery of a specific content/skill area (which is CRT information).

## Minimum Competency Tests

Minimum competency tests are developed to yield criterion-referenced interpretations. That is, the purpose of giving a competency test is to find out what a student knows vis-a-vis prespecified objectives. Test construction procedures should therefore be much the same as those for criterion-referenced tests, although sometimes test developers will select items based on their ability to discriminate between those who pass the test and those who fail. Score interpretations are also similar to those

of criterion-referenced tests, except by definition, scores on minimum competency tests can be interpreted as acceptable or unacceptable performance, i.e., passing the minimum standard or not. (Developers of other criterion-referenced tests also sometimes indicate a passing score or cut-off score indicating mastery.)

Minimum-competency tests: Limitations. Minimum competency tests share the problems and limitations of other standardized tests because the consequences of test performance may be very severe, e.g., students not passing a grade or not receiving a high school diploma, these problems seem more serious.

The task of deciding what goal or objective any test should assess is always difficult. The task of defining objectives that a minimally competent person should attain is complex indeed. For example, what skills are necessary for survival? What skills are truly basic and necessary? For whom? For what? In the absence of emperical data these decisions represent values and opinions which are known to differ among different individuals and groups.

In addition to definitional problems, minimum competency tests, like criterion-referenced tests, in general, suffer from problems inherent in any new technology. The technical quality of some of these tests may be questionable. Reliability and validity are especially important, yet it is not always clear which statistics are best to describe these test properties.

Setting a passing or cut-off score for minimum competency tests is another problem area. Various methods have been advanced to deal with this problem, including the use of expert opinion, past performance of

students on the test, statistical models incorporating the probabilities of misclassification, i.e., identifying a "competent" student as incompetent and vice versa, costs, etc. Although consideration of all these factors will likely yield the most reasonable cut-score, current practice often deviates from this ideal.

In order to help insure student competency, the results of minimum competency tests should have implications for student remediation. As with criterion-referenced tests in general, the number of items assessing each objective often is not sufficient to determine reliably student mastery, and test objectives frequently are not defined well enough to suggest what specific instructional remedies are necessary.

Tests in Native Languages

Achievement testing in languages other than English is constrained by the limited availability of such tests. Reading and math tests are often more available than tests of science and social studies. There are some standardized tests which have parallel versions in Spanish that have been standardized for reading and math. A Guide to Assessment Instruments for Limited English Speaking Students has listings of the tests available in various languages and includes critiques and technical information concerning each tests.

When commercially developed tests are not available, evaluators have several options. Teachers can be requested to monitor achievement on their own teacher-developed tests and the workbook or, if appropriate, you can use tests included in commercially available learning and assessment systems. If the teacher is using a science text in Korean which includes test items, these may be used. An evaluator or teacher can develop a checklist of the

objectives covered in their science curriculum and text, for example, and document the achievement of the objectives for each student. Summatively, the percentage of students across the class who achieved each objective can be documented as a measure of their achievement in science in Korean. An individual checklist of the skills which students have mastered in math and science can become part of a student assessment profile. In this way, subsequent teachers will know which skills have already been learned and thus may plan their curriculum accordingly.

## Technical Quality

### Validity

Validity has to do with an instrument's appropriateness for accomplishing your purposes. These purposes might be to describe people or to make a decision about them, or both. If you expect your audience to accept the results from an instrument used for this purpose, you should demonstrate construct, content, or concurrent validity.

Construct validity refers to how well the instrument measures the underlying trait, such as "reading ability," creativity," or "mathematics facility," that it purports to measure. Demonstrating construct validity demands clear definition of the construct, then presentation of logical arguments, credible opinions, and an accumulation of evidence from correlational or criterion group studies--all aimed at ruling out alternative explanations of the instrument's results. Few, if any standardized achievement tests claim to have construct validity.

Content validity refers to how well the items give appropriate emphasis to the various components of the construct. To show content

validity, you should provide evidence that that instrument contains a set of items which sample the construct's various sub-areas and give each of its proportionate emphasis. In practical terms, content validity refers to how well an instrument represents program objectives. It is the most important type for program evaluation because unless what your students have learned in the program is covered by the test content, the results will be misleading.

Concurrent validity is calculated when you use the results of one measure to predict the results of an alternative contemporaneous measure. Predictive validity justifies a test's usefulness for making decisions about people, such as selection for a special program. A case is usually made for predictive validity by demonstrating that the instrument correlates well with other valid measures of the behavior it is intended to predict.

The credibility of your evaluation depends a lot on the quality of your instrumentation. When reporting results, then, you also should provide evidence to show that you employed valid measures.

Threats to validity. Below are commonly recognized threats to the validity of an instrument. When selecting an instrument and when reporting results, you should take the following potential threats into account.

1) Lack of standardization in test administration. You need to ensure that your test is given in the same way in classroom B as in classroom A. Therefore, the method of giving the test must be carefully controlled and followed exactly by all test users.

2) Too few items per objective. Although the test as a whole may be a valid measure of achievement in the general subject area it represents, you will be unable to make a valid judgment about the presence of the particular subskills in the test unless they are each measured by several items. Rules of thumb for item sufficiently differ, but in many cases, five to eight items per objective are needed in order to make accurate decisions about the extent to which a student has mastered a skill.

3) Tests measure the skill too narrowly. Tests that measure few or narrowly defined skills are not necessarily poor tests, but they are inadequate for making judgments about program effects. Ideally, you will assess student achievement in a broad range of areas that the program is intended to address.

4) Mismatch between the objectives of the test and the kinds of skills actually tested. There is no substitute for reviewing the entire test, from technical manual to items, in order to determine the accuracy with which the items describe the content or objectives to be tested. A test may purport to measure "writing skill" or "reading ability." However, a review of actual items may indicate that the test actually requires students to perform such narrower tasks as proofreading skill or literal comprehension. Although such items may yield information of use in your evaluation, care must be taken that you do not infer that students have acquired skills that are broader in scope than what the test has tapped.

5) Mismatch between objectives of the test and objectives of the program. If you are making decisions about program effects and whether students have achieved program goals, the overriding criterion in test selection will be to find a test that measures program goals. You should review the test to determine whether the items as written reflect the intent of the program goals. You should see also whether there are a sufficient number of items to be able to make accurate decisions about the extent to which goals were attained.

6) Curricular dependence. Test format and/or wording are tied to particular curricular materials. Some tests are developed for a particular curriculum or to assess instruction when particular materials are used. Your students may be unfamiliar with the vocabulary or format of test items if they were not exposed to the curriculum or to the materials. Thus, you should check to see whether the test is based on specialized vocabulary, content, or concepts.

A note about out of level testing. For limited English proficiency students, for whom English is a second language, the level of an achievement test is an important consideration. Since limited English proficiency students are usually underrepresented in the norming sample when a standardized achievement test is developed, the test, strictly speaking, is not a valid indicator of their level of achievement until their mastery of English becomes equivalent to that of the monolingual English speakers who formed the norming sample.

When students are acquiring English as a second language and need to be administered a standardized achievement test, out-of-level testing may be advisable. If students are in fifth grade, then a third grade test may better represent the skills they are learning and thus may provide a better arena for the students to demonstrate the skills they have acquired. For example, their vocabulary would be more like that of the third grade monolingual English speaker than a fifth grade English monolingual student. The important issue in out-of-level testing is curricular relevance.

Guidelines for selecting valid tests. Selection of valid student achievement tests is dependent upon adoption of a systematic decision-making process for test selection. Test validity is a function of matching the information provided by the test with the information needs of the evaluation. The process by which this match is made is twofold. First, you must identify the decisions in the evaluation for which student achievement data are to be used. Second, you will have to ascertain the match between the curriculum and test content. Some of the questions you may wish to ask

while reviewing tests are presented in Figure 3-2. Figure 3-2 summarizes important considerations in selecting tests for your evaluation decisions. Figure 3-2 outlines the major questions related to matching test content to curricular objectives. Both levels of decision making are important in insuring test validity and guarding against the threats to validity described in the previous section.

Reliability

Judgments about test reliability answer the question, "Does the test give consistent results?" A test is reliable if it consistently serves the same function and gives the same information on repeated measures with the same individuals. If scores are subject to random fluctuation, then decisions made on the basis of test results are not generalizable across time and student groups, and have no credibility.

Figure 3-2

Selecting Tests to Match Local Evaluation Purposes

1. What decisions are test results to be used for?

   a. Selection: Scores should be interpretable in terms of 1) the correlation between the obtaining score and the criterion variable or 2) correlation between test score and program success.

      Test items should discriminate between those students who do well on the criterion and those who don't.

   b. Classification: If the test is to be used to classify students as masters/non-masters, scores should be interpretable in terms of what skills the competent student possesses.

      Items should be constructed to yield descriptive information.

   c. Description: If the test is used to diagnose program or student strengths or weaknesses, it should have a clear descriptive scheme that will provide information for remediation and an adequate number of items per objective so that decisions about student performance are stable.

2. To what situations do decisions based on test results generalize?

   a. Are decisions to be made about groups of students or individuals?

      Decisions about individual students generally require a larger number of items per objective on the test than those to be made about groups.

   b. What was the composition of the sample population upon which the test statistics were obtained?

      Tests which are designed to make accurate decisions about heterogeneous samples do not always make accurate discriminations among homogeneous groups or subpopulations such as bilingual students.

   c. Are test statistics, such as reliability and validity coefficients the appropriate ones for the decision context in which the test is to be used?

Split half reliability is a measure of how much the test items elicit the same skill from a single student on a single test occasion. In other words, it shows the internal consistency of a test – the extent to which different items seem to be measuring the same thing. The items on a test (or subtest) are divided into two, for example, odd and even numbered items, and students' performance on the two halves are compared. A coefficient of .60 or better indicates that both halves, when combined together, perform their functions consistently.

Kuder Richardson Approaches, known as KR-20 and KR-21, and alpha also assess the internal consistency of a test or subtest. These kinds of reliability are especially important when you want to use test results to diagnose student or program strengths and weaknesses.

Guidelines for selecting reliable tests. As you have seen from the previous discussion, there are several kinds of test reliability. This puts responsibility on the test consumer for selecting the test with the appropriate kind of reliability for the decision purpose at hand. If, for example, you are selecting a test to measure student achievement over the year, split half reliability--a measure of the degree to which items in the same test are measuring the same thing--does not guarantee that the test is reliable for your purposes. You will need a test that is reliable in terms of two test administrations spaced widely apart. If you are using two forms of the test, you may wish to consider alternate form reliability; if you are using the same form and level, you will want test-retest. Check carefully to see the time period between administrations of the alternate forms. If

they were administered simultaneously, you do know that the two forms are measuring much the same kinds of skill at that time. You do not know, however, if scores on Form A will correlate with scores on Form A-B when the two are administered at different times to the same population of examinees. Since this is the kind of decision you will be making if you are selecting a test for beginning of year pretesting and end of year porttesting, you must be aware of the limitations on your inferences about test reliability.

When making decisions about mastery of objectives or student minimal competency, you may wish to consider decision-consistency of the test. How often are the same examinees classified into the same categories on this test? What kinds of errors are associated with labelling a student as a master or non-master of a certain objective? These criterion-referenced reliabilities are often not reported but may be examined for your own students by constructing simple contingency tables.

Reliability, like validity, is not a fixed characteristic of a test. It is relative to the test purpose and to the examinee population. Since the test statistics of published tests are obtained on student populations where limited English speaking, low socioeconomic, and ethnically different groups constituted only a small percentage of the standardization sample, there is no clear evidence that these validity and reliability statistics reported will be the same for your students. If you have data processing of test scores, you can request that reliability coefficients be computed for your students. If you are using the same test more than once with the same group of examinees, you can have decision consistency, test-retest or alternate form coefficients calculated. In this way you will have better information about score stability and accuracy for your students.

## Practicality

We have described the technical qualities that should influence test selection. Now we turn to practical considerations. Practicality refers to the economy, ease of administration and scoring, and ease of interpretation of the instrument. Unfortunately, this last feature, practicality, often leads to the selection of inappropriate evaluation instruments. There are two aspects of practicality you will need to consider, utility and cost.

## Utility

Administrator qualifications and training. The test to be selected must also be viewed with the potential test administrator in mind. For example, some tests require administration by a psychologist specifically trained to administer and interpret the test. If such tests are necessary, then arrangements must be made to locate appropriately trained test administrators. If not, then tests which can be administered by available trained staff need to be selected. Other criteria for administrators should be considered, such as the administrator's fluency in the student's native language and familiarity with the student's culture. Studies have shown that these factors influence student performance on tests.

Test directions. Some tests are often carefully constructed in terms of the sequencing of sentence structures from simpler to more complex, yet the directions of the test are sometimes of much greater complexity than the material to be tested. One test, for example, requires that the student must listen and comprehend 228 words of directions before responding to the test's single word items. Complex test directions can therefore

be a stumbling block for students. The real test becomes understanding and following the test directions rather than answering the test items. The testing of math skills, for instance, is often embedded in written examples whose directions change with each example. Therefore not only math skills, but comprehension of verbal material, is being assessed. Understanding changing directions is a skill in itself, and teachers might consider devoting teaching time to test-taking skills. A recent study by the Federal Trade Commission has shown that standardized tests, including certain aptitude tests, are vunerable to improvement after coaching in test taking skills.

Test Formats. Some tests for bilingual students have been developed with parallel versions in two languages presented on facing pages of the same test. While this idea seems plausible since a student might not know a word in one language and could refer to the other language, in practice it can be confusing and time-consuming for the student who reassures himself or herself each time of the interpretation of one test passage by reference to the other. Since personality variables may lead to outcomes which do not relfect the achievement of content or skills, the format of this kind of "bilingual" test is not recommended.

Scoring. Tests can be hand or machine scored. In this electronic age, it seems unlikely that any large scale testing program would be undertaken without machine scoring. Although hand scoring saves money, it is expensive in staff time.

In investigating scoring options, you should check out the types of scores available for different tests to make sure the test can provide you the information you need. The following types of scores may be available for norm-referenced tests (the last section describes what each of these scores mean):

1. Raw Score
2. National Percentile Score
3. Local Percentile Score
4. National Stanine Score
5. Local Stanine Score
6. Standard Score
7. Grade or Age Equivalent Scores

For criterion-referenced tests, scores available, in addition to the raw scores are percentage correct, and in some cases, objective mastery scores.

Some publishers offer a variety of reporting options. For example, test results can be provided at the school, classroom or individual student levels. Some even provide test reports for parents. Be aware that not all score reports are equally comprehensible and usable. The organization and format of the reports can facilitate or inhibit the use of the test results. You should therefore check out how scores are reported, particularly if you intend to use the information for diagnostic purposes.

## Cost

The cost of the test obviously is a very important practical consideration. Your basic cost estimate should include the following:

1. Test Coordinator's Handbook
2. Teacher's Guide
3. Examiner's Manual
4. Test Booklets and/or Answer Sheets
5. Number 2 Pencils and Scratch Paper
6. Basic and any additional scoring services and reports desired.

The per pupil costs for standardized tests usually range from 20 cents to 80 cents per pupil, exlusive of scoring and score reports. Computerized scoring is recommended for accuracy and staff morale and usually costs about 90 cents for basic scoring service. In estimating costs, be aware of inflation. Tests may be ordered in July for the upcoming school year, but when the Spring tests are scored, the prices may have increased. Anticipate these increases by adding perhaps a five percent inflation factor.

Some companies also provide practice and/or locator tests for an additional charge. Practice tests, as the name implies, are intended to acquaint students with the format and directions for the test and give them practice in test taking under standardized conditions. Locator tests assist the teacher in determining what level of a test should be administered to each student. Although these tests inflate the testing budget, they can be very helpful to both students and teachers.

Summary

The selection of appropriate student assessment instruments is a complex process that considers three characteristics of test use: validity, reliability and practicality. Each of these screening criteria, in turn, must be employed in the context of the program you are evaluating. You need to know information about the following: What is the program like? What do decision makers wish to know? What kinds of help will you have in collecting data? What kinds of instruments or non-formal assessment devices are already available? This melding of program demands and the constraints on data collection with instrument quality considerations makes the test selection process time-consuming. Nevertheless, the time spent in considering student assessment instruments is a crucial factor in the quality and kinds of information you will be able to provide about the bilingual program.

Figure 3-3 summarizes the contextual variables that affect the kinds of tests you will need. For each of these variables, population characteristics, program information needs, staff expertise, and types of available tests, you will need to ask yourself some questions that can guide the test selection process. Once you have mapped out the constraints on the testing process and are familiar with the context in which student assessment will be carried out, you will then be able to consider the technical and practical merits of available tests. Systematic test selection will help you provide better, more accurate information to the decision-making audiences of the evaluation.

FIGURE 3-3

Questions to Guide Test Selection

I. Population Characteristics

1. How many different language groups will be tested?
2. Which grade levels are to be tested?
3. Are students functioning at grade level or will I need out-of-level tests?
4. How many language skill levels will be tested?

II. Program Information Needs

1. What types of student performance comparisons are required by funding agencies?
2. What student performance information is needed by program staff?
3. What are the program goals and objectives?
4. Which skills best reflect these objectives?

III. Test Administration Staff Expertise

1. Who will administer the tests?
2. Who will score the tests?
3. What special training will staff need to use selected tests?

IV. Types of Tests Available

1. Should I select standardized or non-formal instruments?
2. Do I need a criterion or norm-referenced test?
3. Can special purpose achievement tests be used?
4. What types of language tests should I consider?
5. Which available tests fit program objectives?

Questions to Guide Test Selection

V. Technical Quality

1. What kinds of empirical validity are reported for the test?
2. How can I determine its curricular validity for my program?
3. Is item vocabulary and content appropriate and inoffensive?
4. What kinds of reliability coefficients are reported?
5. Can I request test statistics with score reports to confirm reliability?

VI. Practical Considerations

1. Can an "untrained person" administer the test?
2. How much time does test administration, taking, and scoring consume?
3. Do students understand test directions?
4. Are tests easy to score?
5. Are scores easy to understand?
6. What is the cost per pupil of test administration and scoring?
7. Are tests reusable, can economies of scale be realized with test use?

# Appendix A

## Managing the Collection and Reporting of Test Data

## Managing the Collection of Test Data

### Preparation for Data Collection

If you are responsible for evaluating a large number of classrooms, or more than one school district, the student assessment task becomes one of effective management as well. Although as an evaluator you probably will not select tests for internal program assessment needs, you may need to use these data for your evaluation. Student progress data is invaluable for providing formative information to project staff. You will need to know which students exited from the program, when and why, in order to locate them for posttesting or to describe this population in your evaluation report. Thus one of your first tasks in managing the collection of test data is to identify already existing data that you will need, and find out how to access it.

One commonly used source of internal assessment information is the student cumulative record, developed by the American Guidance Association to record annual achievement test results, attendance data, and some home background data. The card presently used, however, does not include space for the results of the assessment of students in their primary language nor other data which is significant to bilingual programs. There is no provision on the Cumulative Record Card for such data as the language(s) used in the home, the dominant language and the Lau category of the student when initially tested, as well as changes after participation in the program, the results of oral language proficiency assessment in both languages, and assessments of content and skills in the primary language.

The Cumulative Assessment Profile (Saravia-Short, 1979) presented on the following pages, is one strategy for aggregating the various assessment

data for each student. The Profile, or a similar record keeping form, can be kept in each student's folder as an addendum to the Cumulative Record Card.

Developing a Cumulative Assessment Profile (CAP) also can assist you, as evaluator, in planning an efficient testing strategy, one that minimizes the danger of overtesting. The process of outlining the full range of needs may allow you to identify efficient testing alternatives, such as the use of one instrument for multiple purposes. For example, if you are using cohort comparisons or a time series design for your evaluation (please refer to Planning for Evaluation Design, Workshop II for discussion of these designs), the spring posttest data from each previous year could possibly be used as the pretest data for the following year. Such a strategy eliminates testing in the fall, and limits standardized achievement testing to one annual period in spring. By coordinating state-mandated testing with program evaluation, you minimize the possibility of overburdening students and teachers with too much assessment. You may discover other "overlapping" test purposes by working with program staff, exploring their testing needs in conjunction with external evaluation requirements.

Once you have a system for acquiring internal test data, you will need to set up a record keeping system for the test data you intend to collect separately as part of the evaluation design. Several examples of record keeping strategies are presented later in the section in the discussion of "Recording Data for Analysis".

You will also need a brief description of the instrumentation for your evaluation. Such a description might include:

FIGURE 3-4

revision

STUDENT CUMULATIVE ASSESSMENT PROFILE

Information on Student Background and Achievement

A. Student Background Information

(1) Birthdate:

(2) Place of birth:

(3) Parent/Guardians' Birthplace:

(4) Previous school location(s) by grade:

1. _____ 4. _____

2. _____ 5. _____

3. _____ 6. _____

(5) Attendance/number days present per grade:

1. _____ 4. _____

2. _____ 5. _____

3. _____ 6. _____

B. Language Assessment Profile

(1) Language(s) used in the home: _____ _____

(2) Language parent prefers notices, report card sent home: _____

(3) Language(s) used with:

| | English | Other |
|---|---|---|
| Parents: | | |
| Siblings: | | |
| Friends: | | |

(4) Lau Category:

| Lau: | A | B | C | D | E |
|---|---|---|---|---|---|
| Date: | | | | | |

(5) Results of Dominance Test (X=Dominant Language)

| Date | English | Other Language |
|---|---|---|
| | | |
| | | |
| | | |
| | | |

(6) Results of Oral Language Proficiency Tests:

| Date | English | Other Language |
|---|---|---|
| | | |
| | | |
| | | |
| | | |

C. Student Achievement Profile:

Achievement in Primary Language

Reading in Primary Language, by grade

| Grade | Test | Score | Dist. Norm | Grade | Test | Score | Dist. Norm |
|---|---|---|---|---|---|---|---|
| 1 | | | | 4 | | | |
| 2 | | | | 5 | | | |
| 3 | | | | 6 | | | |

Math in Primary Language, by grade

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 1 | | | | 4 | | | |
| 2 | | | | 5 | | | |
| 3 | | | | 6 | | | |

Science in Primary Language

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 1 | | | | 4 | | | |
| 2 | | | | 5 | | | |
| 3 | | | | 6 | | | |

Social Studies in Primary Language

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 1 | | | | 4 | | | |
| 2 | | | | 5 | | | |
| 3 | | | | 6 | | | |

Annual Achievement Testing

English Reading Level by grade

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 1 | | | | 4 | | | |
| 2 | | | | 5 | | | |
| 3 | | | | 6 | | | |

Math Level in English, by grade level

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 1 | | | | 4 | | | |
| 2 | | | | 5 | | | |
| 3 | | | | 6 | | | |

Science in English

| Grade | Test | Score | Dist. Norm | Grade | Test | Score | Dist. Norm |
|---|---|---|---|---|---|---|---|
| 1 | | | | 4 | | | |
| 2 | | | | 5 | | | |
| 3 | | | | 6 | | | |

Social Studies in English

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 1 | | | | 4 | | | |
| 2 | | | | 5 | | | |
| 3 | | | | 6 | | | |

Criterion-Referenced Testing

Math Skills Checklist Month/Year

1.
2.
3.
4., etc.

D. Placement in Instructional Programs

1. Placement in Bilingual Programs

Date: Grade:

2. Reviewed for transfer to English Curriculum:

Date: Outcome:

_____ _____

_____ _____

_____ _____

3. Follow-Up Assessment

Date: Outcome:

_____ _____

_____ _____

_____ _____

4. Referral to Support Service:

Date: Outcome:

_____ _____

_____ _____

_____ _____



Developed by: Marietta Saravia-Shore

- The title and publisher of the test, form number and levels
- Objectives measured
- Pertinent data for the final report: reliability, norms, cut-off scores

A final preparation for data collection is how you plan to have students record their answers to the test. In general, separate answer sheets are most convenient for data collection. They are unsuitable, however, in the following cases:

- The examinees are in grade 3 or below
- The whole measure and all answers can easily be put on one sheet of paper
- There are likely to be more errors resulting from students transposing their choices than from having clerks add item scores across pages in a test booklet.

If a separate answer sheet is used, try to limit its use to one per student even if you have to use the reverse side. This is recommended so that you can avoid confusion and errors that might arise in trying to match separate answer sheets for a given student. If multiple answer sheets are needed, be sure to have a foolproof method of matching them should they become separated.

If a measure contains more than one essay item, it is better to score a given question for all the examinees before going on to score the next question than it is to score all the questions for one examinee at a time. This process is facilitated if you have students put each essay answer on a separate page. Figure 3-5 summarizes some other common practices employed by evaluators in planning for the collection of test data.

Important Dates in Scheduling Data Collection

Once you have a "picture" of which tests will be used, where you will obtain other assessment data, and how student responses will be collected

and scored, you will need to schedule your tests. The "test", of course, may be written or oral, an observation, interview, self-report and so on. In scheduling actual data collection you will need to select testing sites and check with conflicts with other other school events such as fire drills, assemblies or special sessions. You will need to notify the school secretary, teachers involved, and perhaps even other faculty about when testing will take place and where. You should also schedule make-up test dates for absentees.

In setting up test dates you must allow time for the following activities:

- Printing of the tests
- Distribution of tests to teachers who may be administering them
- Training sessions for scorers and test administrators
- Collection of answer sheets and scoring of results.

If you have final reports due in May or June, you may have to begin scheduling posttesting as early as February. Less formal data collection, of course, can be done on a shorter timeline. Nevertheless, it is a natural tendency to underestimate the amount of time necessary to complete data collection activities.

Recording Data for Analysis: General Considerations

Determining what to record. Prior to recording data for analysis you will need to decide what kinds of information needs to be recorded. In general the old principle, "never throw data away", is sound. If you are in doubt about whether to record one or more scores, such as the total test score and the subtest scores, err of the side of recording too much data. Extra information often becomes essential information after initial data analyses are completed or decisions are made to modify the analysis procedures. Then

too, extra information may provide a means for checking possible errors because it often points out discrepancies in your data.

Be sure that you record enough information to ensure the possibility of doing supplementary data analyses that are routinely done for achievement tests. Examples of these analyses include the following:

Item difficulty analysis: This is to determine how difficult or easy items are by computing the proportion of students who have answered each item correctly.

Distractor analysis: This procedure will tell you whether students are selecting certain unexpected distractors. It provides information about whether the distractors are effective, the scoring key is accurate, and the instruction appropriate.

Item discrimination and sensitivity analysis: You may need this information if one of the purposes of the test is to sort students effectively into certain prespecified groups, such as "proficient" and "not proficient".

Reliability analyses: If you wish to report test reliability for your particular group (which may very well differ from the reliability statistics provided by the standardization sample) you will need individual response data for each item in the test as well as data from at least two administrations of the same test, or alternate forms of the same tests.

If you plan to do secondary or supplementary data analyses, you will need to consider the kinds of scores that should be recorded so that you will have the appropriate information for these analyses. There are several ways to report scores: you can report each individual student's response to each item; you may note only whether students answered an item right or wrong; you may wish to record subscores as well as total scores; you should also consider whether you need raw scores or derived scores, and if the latter, which kinds of derived scores. The next section provides some brief guidelines that will help you to decide the kinds of scores you may need for data analyses.

Individual item responses. If you are using computerized test scoring, you will need to program the entire answer sheet into the machine. Mark-sense cards or special answer sheets and pencils save recording this information, but these techniques are cost-effective only when schools have their own scan-tran machines or large-scale testing is involved. Individual responses to each item are also necessary if you wish to analyze the pattern of incorrect answer choices to determine whether the answer key is incorrect or that certain items may have some content "bias" (act differentially for different subgroups of students). If you have this information you will automatically have all possible information for subsequent analyses.

Item scores. You will need to record whether a student got an item correct (1) or incorrect (0) if you are doing any item analyses such as difficulty or discrimination. Item response data is also necessary for computing internal consistency coefficients.

Raw v. derived scores. Raw scores are the basis for all derived scores. If you record only derived scores, then later wish to make comparisons or interpretations that a particular derived score is not designed for, you will have to go back to the raw score again and use the norming tables to obtain the set of scores needed. A safer practice is to record raw scores. If you use computer scoring, you may request the printout to provide these scores as well as a variety of derived scores that will enable you to examine test results from several perspectives.

Scoring Guidelines are summarized in Figure 3-5.

## Scoring Guidelines

Figure 3-5

Hints for scoring selected response tests.

1. Construct a scoring key or guide.

2. Have scorers use a template which contains holes for the students answer choices, or

3. Use a SCAN-TRON or other automatic scoring machine.

Hints for scoring constructed response tests.

1. Score a given question for all examinees before scoring the next question for each examinee.

2. If more than one question per examinee is scored, the papers should be shuffled after one question for all examinees is scored. This helps to reduce biases that might arise when a scorer grades an answer after having just read an especially good or poor answer.

3. Scorer compares answer to a key or guide and makes a decision as to the answer's quality and/or assigns points for gradations of quality in accordance with a scoring guide.

4. Independent scorers judge the quality of each answer and reach consensus, or assign an average rating, without using a standardized key or guide.

5. Periodic checks are necessary to ensure that different individuals scoring the same question are using equal standards of performance for a given score.

   a. In many evaluation studies, it is usually more cost-effective to have a few scorers evaluate several answers by using only a sample of all the answers written than to have all the answers evaluated by only one scorer.

   b. If more than one scorer is used per answer, they should read the answers in a different order.

6. Alternative correct choices should be listed for each completion question.

7. If partial credit is given for a question, the standards for such credit should be specified so that all scorers can use the same standards in the same way.

Hints for correcting common scorer biases and errors.

1. Scan answer sheets before scoring to identify double choices. Template scoring hides these!

2. Periodically have a given set of papers scored twice or by a second scorer to keep scorers alert for clerical errors (such as failing to mark out incorrect answers or improperly counting the number correct).

3. Disguise group membership of each student by using code numbers rather than labeling a given set of papers in a way that would easily indicate membership (in a given school, class, treatment condition, etc.).

4. Provide a standard form for recording scores and/or provide spaces on the answer sheet for subscores and scores.

Recording Systems

The method you use to record data is subject to several considerations. The brief list provided in Figure 3-6 is an example of the factors that influence the method you will choose for recording data. Because of the numerous considerations and unique needs of any given evaluation, it is impossible to provide a description of all the different methods you might use for recording data. You should try to record data as few times as possible; the more often data is transposed, the more opportunity exists for clerical errors. With this general caveat, the remainder of this section will present three methods for recording data: the individual student record form, the classroom data record form, and the program data record form.

Individual student data record form. In this procedure, you will fill out a separate form for each student, similar to the Cumulative Assessment Profile suggested at the beginning of this section. Each form should identify the student by name, class, school, and language proficiency. These forms are especially appropriate when you are doing subgroup analysis, small group testing, and must report on several linguistic groups receiving the same program.

Classroom data record form. This form, like the individual student form, should identify students by name or ID number and should include a list of the measures on which data is to be collected. A separate page may be used to describe the measures and their objectives. Although what you will record depends upon your particular project, some things to record might include:

- Students scores on a measure, subscores, and derived scores.
- Date the data was collected.

Considerations in Selecting a Data Recording System

Figure 3-6

- How many students are involved in the data collection enterprise?
- What are the units of analysis of interest? Individuals? Classes? Schools?
- How will students be grouped for analyses purposes? Language group? ability groups? Sex? Type of bilingual program?
- Which kinds of data analyses will be done?
- Will data be analyzed by hand, calculator, or computer?
- Is all comparable data to be collected at the same time or is it to be recorded when individual students complete certain instruction?
- Who is responsible for recording data, analyzing and interpreting it?
- What other people need access to the data? Teachers? Parents?
- How confidential must data be?

Program data record form. The program data form will record all data on all students in as condensed a form as possible, using numbers to represent people and data. The primary advantage of this approach is that computer analyses may be set up directly from program coding sheets. This method can be used with both large and small projects and is often the second step of data recording for the two methods mentioned above. Since this method is most often associated with computer data analyses and keypunching, and since the steps are the same in both methods, we will present the method for reporting program data in a step-by-step outline following. These steps are presented in Figure 3-7.

## Recording Program Data

Figure 3-7

Step #1: Assign an identification number to each student.

a. The kinds of information that go into an ID # include:

1. sex
2. student's name and/or a number representing it
3. birth date
4. special subgroup, e.g., experimental or control
5. location--school, class, district, etc.
6. testing date--e.g., pretest or posttest, etc.
7. student's grade level

b. Some or all of this information may be recorded on the students' answer sheets prior to the testing, especially with large projects; if this is not done, special care should be made to ensure that a given set of data can be identified as coming from a given student.

1. One way to facilitate this is to provide each test administrator (usually a teacher) with a premarked envelope for each group (or class) of students. All the data from this group are then placed in the envelope.

2. Extra special care must be used in those cases where there is a single measure given twice, such as in the case of a pre- and posttest.

c. The ID # should be placed on all the separate documents (test and answer sheets, questionnaires, etc.) for each student. If students record some or all of the identifying information, it should be double-checked for accuracy --especially with younger students where it may even be necessary to have the teacher record it.

Step #2: Assign code numbers to each kind of identifying information in a systematic fashion. For example, if you wanted to record the student's name, sex, and related information, you might do it as follows:

a. Project # - if you are doing more than one project, you might wish to number them such as:

1 = K-3 reading
2 = Title I mathematics
3 = Junior high sex education

b. School and class

11 = 1st period at Eastchester J.H.S.
12 = 2nd period at Eastchester J.H.S.
.
.
.
21 = 1st period at New Briar J.H.S.
22 = 2nd period at New Briar J.H.S.

c. Treatment group

1 = Experimental program
2 = Regular program (control)

d. Track

1 = low
2 = middle
3 = high

e. Sex

1 = male
2 = female

f. Racial-ethnic group (when needed for analysis)

1 = American Indian
2 = Black
3 = Spanish surname
4 = Asian-American
5 = White, other than Spanish surname
6 = Other, please specify

g. Student number

There are several ways of doing this, but the simplest and safest is to make a list of all the students in the program and assign them numbers starting from 1. If new students enter the program or if students switch classes, this method is capable of handling such problems.

For example: Aaron, A = 001
Adams, T = 002
Anton, D = 003

Thus, Donald Anton, who is a white middle-track student in the 4th period experimental sex education program at Eastchester J.H.S. would have the following ID #: "3/14/2/1/5/003".

Note: It is not necessary, but it is usually more efficient, to assign numbers to students whose names have been placed in alphabetical order (if not for the whole program, at least by class).

Step #3: Record the data

a. Prepare a table in which each row represents a single student

b. Each column in this table contains a different kind of information. For example, the first set of columns would probably contain the student's ID # and then other columns might contain his test scores, responses to questionnaire items, etc.

   1. The only exception to this rule would occur where only class or group data were known and the analyses involved an examination of several of these classes or groups. Under these circumstances, the rows would contain the scores for each class or group.

   2. Sometimes it is helpful to record the students' names such as when teachers will use the layout sheet for recording additional information about the student during the year, i.e., fill it in as they go.

c. The data can be placed in this table by hand (such as recording each student's score on each test) and/or by keypunching directly from an answer sheet, test booklet, or questionnaire.

d. If keypunching is to be done for computer analyses, then there is a special set of rules governing the construction of this table. These rules are as follows:

   1. The table should contain no more than 80 columns and contain only one number per column. If it might take up to three numbers to record a test score, then three separate columns must be set aside for this purpose.

   2. If it would take more than 80 columns to list all the information on each student, then a second, third, fourth, or even more series of 80 column tables would be needed, since a given card only has 80 columns.

      (a) When more than one set of tables is needed, it is necessary to include the student's ID # in the same place in each separate set.

      (b) The reason for this is that there is a separate card for each student for each set of data. The only way to tie the separate cards together for a given student is through having the same ID #.

Step #4: Prepare a "Card Layout Sheet" for each set of data.

a. This sheet explains how and what information gets punched into what columns.

b. Several copies should be made of this sheet, since its loss would be like trying to use a road map without a legend, names of towns, highway numbers, etc.

(An example of a card layout sheet appears below.

c. Keypunchers work fastest from a roster sheet (usually about 40 rows per sheet with, of course, a maximum of 80 columns).

d. Keypunchers can work directly from an answer sheet or test booklet; but all the information that they should punch, and into what columns the information should be placed, should be specified clearly on a sample. Do not ask keypunchers to try to score the tests as they go, or convert words or letters to numbers, etc. They should just have to "punch 'em as they see 'em".

e. "Verifying" means having a second keypuncher check the first one's work. A special machine is used for this purpose and it indicates whenever there is discrepancy between the two keypunchers. These discrepancies are then checked and the corrections made. You can tell when a deck of cards has been verified by the fact that there is a small groove cut in the right-hand side of the card. (If the groove is at the top of a card it means that a discrepancy was noted between the two keypunchers and this card should not be in your deck as is unless it is checked again.)

f. "Interpreting" means having the numbers printed at the top of the card that correspond with the holes that are punched. This is helpful if you have to find certain cards in your total deck, e.g., those students who do not have complete sets of scores.

g. The card below has been punched, verified, and interpreted.

Step #5:

a. Make a printout of the cards. This is done by feeding the cards into a computer that "reads them" and produces a printed sheet with all the information that has been punched into each card.

b. Duplicate your deck of cards at least once and store the duplicate deck in a safe place (and not near the deck with which you intend to work).

## Reporting Test Results

How you will report test results depends upon two general factors: what test scores are computed and what inferences about student performance you can make from these scores. Each type of test has different kinds of allowable interpretations. We have seen that score interpretations differ between norm and criterion referenced tests. We have also seen that the credibility of these inferences depends in part on whether the test is standarized or is a non-formal measuring device. The subject of this section is not how to interpret and report scores from the many kinds of data collection strategies that you will use in your evaluation, but to focus solely on one type of reporting strategy, that associated with standardized, norm-referenced tests for reporting purposes. Since norm-referenced test data is so familiar to school district personnel and is often requested by school boards or funding agencies, strategies for interpreting and reporting these data will be of interest to a majority of workshop participants.

Interpretation of standardized, norm-referenced test scores is influenced by three factors:

1) The form in which scores are reported.
2) The standardization group (norm group) to whom test scores refer.
3) The test content and other factors measured by the test.

Figure 3-8

Sample Printout: District Level Norm-Referenced Test Score Report

District
Grade 9.6
Level/Form: 16C, 17C, 18C (Levels)

Reading Test
Vocabulary C (Form)

| | | |
|---|---|---|
| N COUNT | 1328 | |
| MEAN GE | 8.4 | Median Grade Equivalent |
| MEDIAN GE | 8.1 | Median Grade Equivalent |
| STD DEV GE | 2.7 | (Standard Deviation Grade Equivalent) |
| MEAN SS | 541.5 | (Mean Scale Score) |
| MEDIAN SS | 535.3 | (Median Scale Score) |
| STD DEV SS | 74.1 | (Standard Deviation Scale Score) |
| MEAN NCE | 42.9 | (Mean Normal Curve Equivalent) |
| MEDIAN NCE | 41.3 | (Median Normal Curve Equivalent) |
| STD DEV NCE | 18.5 | (Standard Deviation, NCE) |
| MEDIAN NP | 34.3 | (Median National Percentile rank) |
| 75%ILE (Q-3) | | |
| GE | 10.3 | DISTRICT-NORMS. |
| SS | 587.4 | Scores that are |
| NCE | 54.4 | equilvalent to the |
| NP | 53.4 | 75-50-25%ile rank |
| 50%ILE (Q-2) | | when only the students |
| GE | 8.1 | in the school district |
| SS | 535.3 | are used for ranking |
| NCE | 41.3 | purposes. |
| NP | 34.3 | |
| 25%ILE (Q-1) | | |
| GE | 6.4 | |
| SS | 490.7 | |
| NCE | 30.1 | |
| NP | 17.1 | |
| N/QUARTER | | |
| 76-99 | 159 | NATIONAL-NORMS: |
| 51-75 | 285 | Number of students falling |
| 26-50 | 373 | into national percentile |
| 1-25 | 511 | rank. |
| %/QUARTER | | |
| 76-99 | 12.0 | Percent of students |
| 51-75 | 21.5 | falling into national |
| 26-50 | 28.1 | percentile ranks |
| 1-25 | 38.5 | |

Accordingly, this section is organized into three parts: A discussion of the printout reporting the scores; a discussion of the characteristics of the standardization sample to whom our students are compared; A discussion of what the test measures and how accurately. At the end of this section is a list of things to remember in reporting standardized, norm-reference scores.

What do NRT Printouts Tell Us?

Look at the sample printout in Figure 3-8. The printout provides the following kinds of information: name of school, number of students taking the test, grade level of students, and test forms used.
The printout also provides group performance data for each school in the district. These group averages are presented with three different kinds of scores: grade equivalents (GE); scale scores (SS); and normal curve equivalents (NCE). Each of these is described in the next section. For each type of score, we also are told one groups' mean score, median score, and the standard deviation. The mean score is the arithmetic average. All scores of all students in a school are added. Then this total is divided by the total number of scores. The mean is a fairly good indicator of "average" performance unless there is a very large range from the top score to the bottom score, or large numbers of students score either very high or very low. When "skewed" (distorted) score distributions occur, we use the median to represent "average" performance. The median score is that point at which 1/2 of the students score above it and 1/2 score below. The final score statistic reported is the standard deviation.

The standard deviation is a measure of the score spread. Small standard deviations indicate that student scores were clustered near the middle of the score distribution; large standard deviations indicate that the range from the lowest to the highest score was very wide.

The printout provides comparative information for the school district in two forms; district norms and national norms. The district norms, reported in the section immediately after the mean score reports for the national percentiles, indicate which actual grade equivalent, scale score, normal curve equivalent, and national percentile score were equal to a district percentile ranking of 75, 50 and 25. Thus, for this district, a student who scored 10.3 grade equivalent in reading, would be ranked at the 75th%tile as compared to his/her peers. This score is better than 74% of the scores of students in this district taking the test and worse than 24% of their scores. The same interpretation holds for the other percentile intervals.

Finally, this one printout displays the number and percentage of students who fall into four percentile ranks: 1-25%ile; 26-50%ile; 51-75%ile; 76+%ile. These are also called "quartiles". Quartile ranks tell us what percent of the students fall into the lower, middle, and upper score ranges when compared with the national norm group.

What do the Scores Mean?

NRT scores are reported in converted form. The raw score (total number correct) for each student is changed to a derived score, either by using a norm table or by a simple linear transformation (multiply by a number, add a number, and divide by the standard deviation). It is the

converted scores that we use to make test interpretations. Each kind of derived score has different interpretations; each also has certain limitations. The following is a brief description of recommended and not recommended test score reports.

The recommended scores have in common the following characteristics:

1) They have straightforward interpretations.
2) They have numerical properties that allow you to add them, subtract them and make direct comparisons among subtests or groups of students for evaluation purposes.

Scores that we recommend you avoid for evaluation reporting are not recommended because:

1) They are easily misinterpreted.
2) They cannot be used to make direct comparisons among groups of bilingual students tested at different times or on different subtests.
3) They are not based on "equal interval" scale so that you cannot use them in calculating gains.

Recommended Scores

Scale scores. Scale scores are linear conversions of raw scores. They range from 000-999. A student with a scale score of 420 has scored 20 points higher than one with a score of 400. Scale scores do not make comparisons to norm group performance like percentiles, GE's and NCE's. Scale scores are independent of test level, form, grade and test date. Because of these qualities, they are used to compare students in different grade levels or to chart the performance of a single group of students over

time or across different subtests. This kind of comparison is not possible with raw scores because these scores have different meanings, depending on the number of items in a test or the score distribution of students taking the test!

Normal curve equivalents: Normal curve equivalents are scores that have been linearly transformed from percentiles. The transformation is to make them into an equal interval scale so that the distance between two NCE's is the same. They have the same meaning as percentiles; NCE's provide a rank score for a student as compared to the national norm group.

Because of the transformation, NCE's can be added, subtracted, and compared. An NCE score of 67 on one test is actually 7 points higher than an NCE score of 60 of another. The primary purpose of using NCE's is to make comparisons across different standardized tests such as those used in the Anchor study. These comparisons assure equivalence of content as well.

Scores Not Recommended

Percentiles. Percentile scores tell how a student's raw score compares with the score distribution in the norm group. Percentiles are obtained by taking the raw score and looking up in a conversion table the rank that this raw score has as compared to the norm group. Thus, if the raw score is 12, and we see on the Conversion Table that 12=80th%, we know that the person who got 12 correct on the original standardization sample, scored better than 79 percent of the students and worse than 19 percent of the students in the norm group. Percentiles compare our students' performance with the ranks of the norm sample.

Percentiles do not tell what percentage of test items were answered correctly. For example, the raw score of 17 on a reading vocabulary test may convert to a percentile rank of 80. Seventeen correct, however, may actually be only 57 percent of the items right.

Percentiles are not equal interval scales. You cannot add or subtract them, nor use them to make comparisons across groups or tests. Raw scores tend to be clustered together near the middle of the distribution, 40-60 percentile, and spread apart near the ends, 1-20 percentile, and 80-99 percentile. As a result, if you score near the middle, you may only have to get several more items right to raise your percentile score quite dramatically.

Grade equivalents. Grade equivalent scores represent the grade and month in school of students in the norm group obtaining particular raw scores. If, for example, a student gets a raw score of 20 that converts into a grade equivalent of 5.6, it means that a group of fifth graders in the sixth month of school averaged 20 items correct on the same test.

Since most tests are administered only twice to the standardization sample, once in November and once in May, all GE scores for other times are estimates. The assumption behind the estimation is that student growth in the subject is uniform during the year; i.e., this growth occurs at the same rate between November and May.

The distance between GE scores are equal for a particular test. However, the GE scores for different subtests of the CAT have different intervals. Thus, 8.2 on the Reading Vocabulary test does not represent the same number/percent of items correct as an 8.2 on the Language Mechanics Test. GE's do not mean a student has mastered skills representative

of that grade level. Also, a student may get a high GE score, such as 8.2, on a low test level, but when switched to a higher grade level test, the score may drop. Some test publishers recommend that grade equivalent scores not be reported to parents because they are so often misinterpreted.

## How Do I Interpret These Scores?

Norm group interpretations. Percentiles, grade equivalents, and normal curve equivalents tell us about how well students are doing when compared to the performance of the national standardization sample. The more unlike the norm group they are, the less appropriate are these comparisons. The Technical Manual for each test has extensive norm group descriptive statistics. The following norm group characteristics are provided by one popular test:

> There were 200,000 students in the original sample. School districts with 11+ pupils were sampled. The group was about 15% black, 8.1% Hispanic, and 7.2% ESL. Only about 17,000 students were sampled from the western states which includes from the Rockies to Alaska and Hawaii.

It is important to note norm group characteristics and that you can decide how valid comparisons between your students and the norm group are. The more your students differ from the standardization sample, the more you will need to decide which kinds of comparisons are informative and which are irrelevant.

## Curricular Implications of Scores

Group achievement of curricular objectives can be investigated by examining the actual objectives of the test and the number of items sampling each objective. Content emphases are reflected in the number

of items per each objective. The more items per objective, the more the test score reflects that particular content. For example, one popular NRT had 238 items. About 85 (35%) of these items are math related, 20% test reading, and 26% test grammar and punctuation. The total score for this particular test, then, is more illustrative of the English curriculum, for example, than the math. Each of the reading, language and mathematics subtests have content differences where some objectives receive more weight in the subtest score than others.

Score accuracy. A final aspect of score interpretation is the accuracy of the score for the particular group being tested. This is a function of test reliability; that is the stability of the test scores over time. Whether group means and standard deviations change greatly between test administrations is both a function of the quality of the test and the age of the students. For younger students, whose growth is ofter rapid, test scores may fluctuate greatly from one month to the next. As students get older, scores tend to stabilize. The reliability of a standardized test varies both by subtest and test level. Table 3-9 shows sections of the California Achievement Test, Levels 16-18C, used with upper elementary and junior high school students.

Table 3-9

Fall-Spring Reliability for Selected CAT subtests

| Subtest: | Total Reading | Total Language | Total Math | Total Battery |
|---|---|---|---|---|
| Test Level | | | | |
| Level 16C | .89 | .84 | .87 | .93 |
| Level 17C | .89 | .87 | .87 | .93 |
| Level 18C | .88 | .86 | .88 | .93 |

These reliabilities will be different for each school district because students' propensity to guess, to leave questions blank, or to learn quickly may differ from the norm sample. Basically, the above coefficients indicate that total test scores for the norm group will be relatively the same after six months of instruction.

A second aspect of score accuracy has to do with how much error of measurement is associated with any particular score. The more error, the less certain we are that a particular score represents group performance. The accuracy of any type of test score reported following is not pinpoint. Rather, test scores should be thought of as probability estimates of performance occurring within a performance band. The size of this band depends upo the amount of measurement error incurred on any one test occasion.

We can calculate the amount of error associated with test scores by having the standard deviation of our sample and their reliability coefficient. If we have no reliability information on our students, we have to use the reliability estimates from the norm sample as an estimate of how accurate our scores are. For each score, we can construct an interval for which we can say "68% of the time, if I were to give this

test over and over again, actual students scores would fall within this interval." An example of the accuracy of one score from the printout follows:

1. The reading Vocabulary Scale Score for 9.6 graders is 541.5.
2. The test-retest reliability for this test with ninth graders on Level 18C of this test is .83.
3. The amount of spread in our students' scores on this test is the standard deviation which equals 74.1.
4. The amount of measurement error can then be calculated by the following formula: Measurement Error=Standard deviation $\sqrt{1\text{-reliability coefficient}}$.

   Measurement Error = $74.1\sqrt{1-.83}$

   Measurement Error = 30.5
5. If we add this 30.5 point error to the mean scale score for reading vocabulary, we get the upper bound of the confidence interval. We then subtract it from 541.5 to get the lower bound. Result: The ninth grade students' scores on Reading Vocabulary, with Level 18C, can be expected to fall somewhere in the range of 510.5 to 581.5 about 68% of the time we give this test to the same group. Different students, different tests, different levels, will change the score range and the accuracy of these scores. This procedure can be completed for each score reported.

The important thing to remember is that scores are not accurate predictors but rough estimates of performance that fall within certain ranges of scores.

APPENDIX B

## Thirty-three Features to Consider When Selecting a Test For a Bilingual Program

This listing is intended as an aid in choosing features on which to compare tests. It may be an aid for the step of rating the importance of test features as well. We suggest having your local experts on each test feature evaluate that feature, so that this listing is not meant to teach you from scratch how to evaluate features that you are not already well acquainted with.

A. Features that are particularly relevant to bilingual programs

Sources of information on the feature: Test booklets, directions to pupils, answer materials, and directions for scoring.

Preferred judges: Educators who are from the pupils' primary culture and, for immigrant pupil groups, are familiar with instructional practices and materials in the pupils' native schools.

Relevant test uses: all

1. Surface fairness: Are different racial, national, or cultural groups portrayed, in words and pictures, representatively and positively?

2. Vocabulary of test items: Are the items either in the test takers' idiom, or else in vocabulary that does not have different meanings for the different groups that will be tested. To be avoided are words that have different meanings in the different dialects of children to be tested (e.g., Puerto Rican, Cuban, and Mexican) and English words whose dominant meaning is different from the dominant meaning of the similar word in the primary language (e.g., English - embarrassed vs. Spanish - embarazada, meaning pregnant); English - actual vs. Spanish - actual, meaning current or present).

3. Item content and response behaviors: Are the contents of the items and the behaviors required for answering consistent with the information and habits of the pupils' cultures and geographical setting?

4. Cultural aspects of visual design: Are test items and answer materials positioned and spaced in a manner that the different cultures to be tested are used to? Negative example: Children who are used to doing division problems in linear or working form are given a placement test with division problems in radical form. Or, arithmetic problems do not have the work space located where the children are accustomed to.

5. Testing procedures: Do procedural aspects such as speeded testing, multiple choice formats, separate answer sheets, etc., make it harder for one of your pupil groups to perform on the test?

6. Objectivity of scoring: Are the guidelines for scoring so clear and easy to apply that different scorers a) know what to do, and b) get the same results with the same pupil responses? The major problem area is oral language production.

7. Guidelines for decision making: Does the publisher give usable advice on how to make decisions about individual pupils on the basis of a combination of information sources, one of them being test scores? Home language surveys and primary language proficiency are two major sources for making decisions about English language proficiency.

This test feature is especially relevant in determining pupils eligibility for a bilingual program and their readiness to be taught in English.

X. Appropriateness of reference groups: Are the groups of pupils who participated in the test's developmental trials, norming, and validating likely to make the test suitable for use with your pupils? Covered below in C.4, D.1, D.4, and G.5.

B. Objectives or domains that a test includes

Sources of information on the feature: Listing in a test manual or continuum chart of the objectives covered at each level of the test.

Preferred judges: Curriculum specialists in the content area of the test. For #4 below, testing and evaluation specialists.

Relevant test uses: Program planning, diagnosis, progress monitoring, proficiency testing.

1. Clarity of objectives: Does the statement of each objective or skill make it clear just what content and behavior are being tested, or would many different types of content and behavior be consistent with each objective? Objectives such as the following are not clear: reading comprehension, critical thinking, arithmetic operations, arithmetic applications. The following objective approaches clarity, but still has much leeway:

> Given a story of 4-5 lines at a fourth grade reading level, pupils will select the main idea. The three distractors will deal with particulars of the story or with generalizations from single particulars.

2. Rationale for objectives: The primary justification for including an objective on a test is that the objective is actually taught in the program classrooms. If a test is being selected to measure the direct effects of instruction, this test feature may be omitted here because it is covered in detail later in the workshop. But if a test is being used for some predictive purpose, such as to test survival skills in reading and math, the choice of objectives should be well justified.

3. Flexibility in selecting objectives: Is an adequate range or number of different objectives covered? For testing at several levels, are the more important core objectives covered at several test levels with items that are appropriate for the respective ages? For continuous progress monitoring (end of unit testing), are single objectives easy to test separately?

4. Number of items per objective: Is the number of test items for each objective appropriate for the intended use of the test. When test results are to be used to make decisions about the specific skills of individual pupils, there must be at least several items per specific skill. For end of unit testing there should be at least eight or ten, and often much more, depending on the desired level of proficiency. When there are only one or two items per objective, then the results are likely to be misleading, because the score for an objective can be greatly affected by carelessness, guessing, or using a neighbor's answer. If a test is to be used to survey a program, not to support classroom instruction, then the number of items per objective can be small, and the variety of objectives should be great.

C. Adequacy of the test development process

Sources of information on the feature: Technical reports, technical manuals, or technical sections of the test manual.

Preferred judges: Testing specialists and evaluators, except for #3 which may also be judged by specialists in the relevant area of curriculum.

Relevant test uses: all

1. Item review: Were the test items adequately reviewed for clarity, reading level, flaws in item construction, etc., not only by reviewers, but also on the basis of pilot testing?

2. Congruency: Is there convincing evidence that the test items measure the skill and content described by their respective objectives? If there are only sketchy objectives, then there cannot be any convincing evidence. Important for objectives-based testing.

3. Representativeness: Is each set of items which gets scored a typical, representative sample of the skill? Or is it instead a biased, untypical sample?

4. Pilot testing: Was pilot testing of the test items carried out, and if so, was it on a representative and sizable sample of test takers?

D. Validation by field testing

Evidence: Technical reports or technical manuals.

Preferred judges: Evaluators and measurement specialists.

Relevant test uses: All, except where noted.

1. Pupils in the field testing sample: Was the production version of the test validated through field testing, and if so, was the group of students in the field test sizable and either representive of the nation or of the groups to be tested in your program? If not, then any data for D.2 - 5 below are not meaningful.

2. Sensitivity to learning: Is evidence reported that instruction is followed by dependable increases in scores on relevant items? The evidence should be free of the usual problems in measuring gain (e.g.: The increase in test scores should not be attributable merely to maturation of the pupils).

3. Consistency of scores: Is appropriate evidence for the reliability of test scores reported? For tests where human judgment is heavily involved in the scoring (e.g.,tests of speaking or writing), it is essential to have evidence of inter-judge or inter-rater reliability. If alternate forms of a test will be used, consistency in their scores for individual test takers should be shown. When scores on individual objectives or subtests will be used (e.g., for diagnosis), then the reliability data should be for such scores, not merely on total test scores. Finally, if the scale of test scores is to be split into two or three categories (e.g., pass/fail; NES/LES/FES), the reliability data should be for that type of decision and not for total test scores.

4. Lack of bias in item statistics: Does the test publisher report statistical evidence that the test functions essentially the same for the different pupil groups that you will be testing? A showing that one group performs somewhat lower overall than another is not sufficient evidence, because that result will occur when the one group has not learned the target skills as thoroughly. Sound evidence will consist of data showing the pattern of item difficulties (that is, the difficulty of each item on a test relative to the difficulties of the other items) is the same for the various groups being tested, or that items cluster similarly.

5. Validity of passing scores: If cutting scores are to be used, is evidence offered that pass/fail scores strongly predict a valid indicator of success?

E. Appropriateness for examinees

Sources of information on the feature: Examiner's manual, test materials, answer materials

Preferred judges: Teachers

Relevant test uses: All

1. Directions to test takers: Are the directions for each subtest clear

and complete? Are there separate directions for each group of items that needs to be introduced separately? Are sample (i.e., practice) items given where needed? The difficulty of the directions is relevant here.

2. Testing time: Is the time required for testing appropriate for pupils like yours? If there are time limits, are they appropriate?

3. Layout, print, and illustrations: Do the visible characteristics of the materials make it easy for the pupil to do the test? Are the print and illustrations clear and large enough? Is the content of the illustrations familiar to the pupils? Is the material for each test item adequately separated from the material for other items?

4. Item difficulty: Covered in a different workshop unit on rating the program relevance of a test.

F. Procedural features

Sources of information on the feature: Publishers' catalogs, examiners manuals, other sample materials.

Preferred judges: As indicated below.

Relevant test uses: All, except where noted.

1. Qualifications of test administrator: Are the qualifications of the test administrator made clear, and are your personnel qualified? Have test givers or testing specialists judge this.

2. Directions to the tester: Are the directions to the tester clear, complete, and easy to use? To be judged by testers.

3. Ease, speed, and flexibility of scoring: Are the desired options for scoring available? For classroom use of scores, either hand scoring by means of a template (or other objective and efficient key) or machine scoring with very rapid turnaround is needed. For program planning, accountability, or program evaluation, slower scoring may be adequate. To be judged in part by the scorers and in part by the consumers of the test scores.

4. Curriculum indexing: Does the test publisher offer an optional index relating the specific skills on the test to specific learning activities or to the lessons and exercises in several series of appropriate instructional materials? Relevant when test scores are to be used to support instructional decisions in the classroom. Appropriate judges: curriculum coordinators and teachers.

5. Availability of alternate test forms: Are two or more parallel forms available? Important when pupils are to be tested more than once, as in pre- and post-testing.

G. Reporting and interpreting test scores

Sources of information on the feature: Technical manuals, examiners' manuals

Preferred judges: The various consumers of the scores

Relevant test uses: All, except as indicated.

1. Choice of score reports: Are the levels of score reports that you want available? Scores are needed for individual pupils on individual objectives for diagnosis, prescription, and ongoing progress verification. Classroom scores for single objectives are useful to teachers for instructional planning. Class, grade, building, and program level scores may be useful for accountability, needs assessment, or program evaluation.

2. Choice of score types: Are the types of scores that you want available? Mastery (pass/fail), domain (% correct), and percentile norms are possibilities. Grade level equivalents should not be used at all. Are the scores given in a form that is easy to use and interpret?

3. Score report or record forms: Are the forms for reporting or recording scores easy to use and appropriate for the test use?

4. Cautions: Does the publisher give information on the limits to interpreting test scores? Information on the amount and sources of error in measurement is useful, as is information on the types and probabilities of decision error.

5. Appropriateness of norm groups: Are the pupil populations for norming the test meaningful to compare with your pupils?